

<https://helda.helsinki.fi>

---

## Transcript expression-aware annotation improves rare variant interpretation

### Genome Aggregation Database Prod T

2020-05-28

---

Genome Aggregation Database Prod T , Genome Aggregation Database Consor ,  
Cummings , B B , Karczewski , K J , Kosmicki , J A , Karjalainen , J , Daly , M J , MacArthur ,  
D G , Färkkilä , M , Holm , M M , Kallela , M , Kaprio , J , Palotie , A , Tuomi , T , Wessman ,  
M & Ripatti , S 2020 , ' Transcript expression-aware annotation improves rare variant  
interpretation ' , Nature , vol. 581 , no. 7809 , pp. 452-+ . <https://doi.org/10.1038/s41586-020-2329-2>

---

<http://hdl.handle.net/10138/325151>

<https://doi.org/10.1038/s41586-020-2329-2>

---

cc\_by

publishedVersion

---

*Downloaded from Helda, University of Helsinki institutional repository.*

*This is an electronic reprint of the original article.*

*This reprint may differ from the original in pagination and typographic detail.*

*Please cite the original version.*

# Transcript expression-aware annotation improves rare variant interpretation

<https://doi.org/10.1038/s41586-020-2329-2>

Received: 21 February 2019

Accepted: 23 April 2020

Published online: 27 May 2020

Open access

 Check for updates

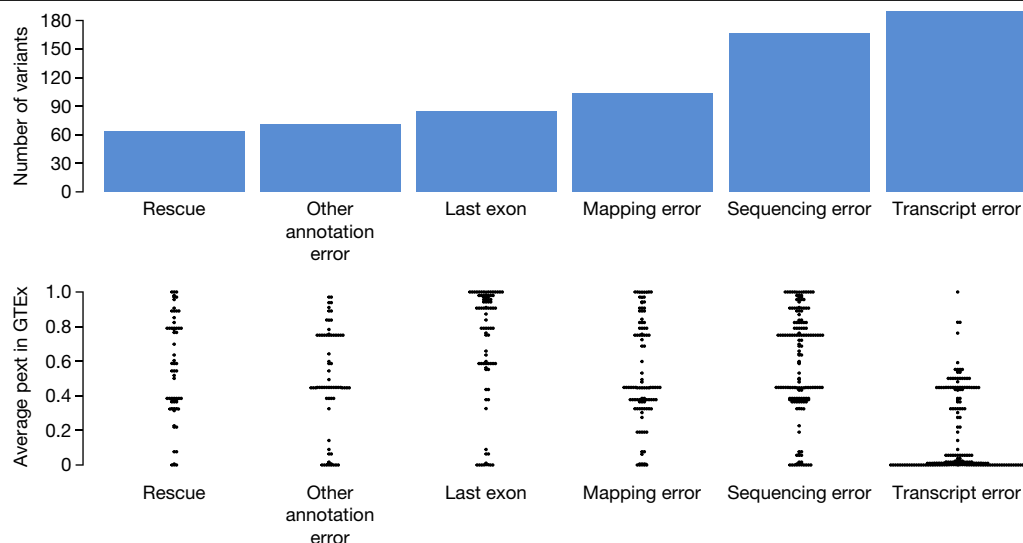
Beryl B. Cummings<sup>1,2,3</sup>, Konrad J. Karczewski<sup>1,2</sup>, Jack A. Kosmicki<sup>1,2,4</sup>, Eleanor G. Seaby<sup>1,2,5</sup>, Nicholas A. Watts<sup>1,2</sup>, Moriel Singer-Berk<sup>1</sup>, Jonathan M. Mudge<sup>6</sup>, Juha Karjalainen<sup>1,2,7</sup>, F. Kyle Satterstrom<sup>1,2,7</sup>, Anne H. O'Donnell-Luria<sup>1,8,9</sup>, Timothy Poterba<sup>1,2,7</sup>, Cotton Seed<sup>2,7</sup>, Matthew Solomonson<sup>1,2</sup>, Jessica Alföldi<sup>1,2</sup>, Genome Aggregation Database Production Team\*, Genome Aggregation Database Consortium\*, Mark J. Daly<sup>1,2</sup> & Daniel G. MacArthur<sup>1,2,145,146</sup>✉

The acceleration of DNA sequencing in samples from patients and population studies has resulted in extensive catalogues of human genetic variation, but the interpretation of rare genetic variants remains problematic. A notable example of this challenge is the existence of disruptive variants in dosage-sensitive disease genes, even in apparently healthy individuals. Here, by manual curation of putative loss-of-function (pLoF) variants in haploinsufficient disease genes in the Genome Aggregation Database (gnomAD)<sup>1</sup>, we show that one explanation for this paradox involves alternative splicing of mRNA, which allows exons of a gene to be expressed at varying levels across different cell types. Currently, no existing annotation tool systematically incorporates information about exon expression into the interpretation of variants. We develop a transcript-level annotation metric known as the 'proportion expressed across transcripts', which quantifies isoform expression for variants. We calculate this metric using 11,706 tissue samples from the Genotype Tissue Expression (GTEx) project<sup>2</sup> and show that it can differentiate between weakly and highly evolutionarily conserved exons, a proxy for functional importance. We demonstrate that expression-based annotation selectively filters 22.8% of falsely annotated pLoF variants found in haploinsufficient disease genes in gnomAD, while removing less than 4% of high-confidence pathogenic variants in the same genes. Finally, we apply our expression filter to the analysis of de novo variants in patients with autism spectrum disorder and intellectual disability or developmental disorders to show that pLoF variants in weakly expressed regions have similar effect sizes to those of synonymous variants, whereas pLoF variants in highly expressed exons are most strongly enriched among cases. Our annotation is fast, flexible and generalizable, making it possible for any variant file to be annotated with any isoform expression dataset, and will be valuable for the genetic diagnosis of rare diseases, the analysis of rare variant burden in complex disorders, and the curation and prioritization of variants in recall-by-genotype studies.

A primary challenge in the use of genome and exome sequencing to predict human phenotypes is that our capacity to identify genetic variation exceeds our ability to interpret their functional impact<sup>3,4</sup>. One underappreciated source of variability for variant interpretation involves differences in alternative mRNA splicing, which enables exons to be expressed at different levels across tissues. These expression differences mean that variants in different regions of a gene can have

different phenotypic outcomes depending on the isoforms they affect. For example, variants that occur in an exon differentially included in two isoforms of *CACNA1C* with diverse patterns of tissue expression result in distinct types of Timothy syndrome<sup>5</sup>. Pathogenic variants in the isoform that exhibits multi-tissue expression result in a multi-system disorder<sup>5–7</sup>, whereas those on the isoform predominantly expressed in the heart result in more severe and specific cardiac defects<sup>8</sup>. In addition,

<sup>1</sup>Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA. <sup>2</sup>Analytical and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA, USA. <sup>3</sup>Program in Biological and Biomedical Sciences, Harvard Medical School, Boston, MA, USA. <sup>4</sup>Program in Bioinformatics and Integrative Genomics, Harvard Medical School, Boston, MA, USA. <sup>5</sup>Genomic Informatics Group, University Hospital Southampton, Southampton, UK. <sup>6</sup>European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge, UK. <sup>7</sup>Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, MA, USA. <sup>8</sup>Division of Genetics and Genomics, Boston Children's Hospital, Boston, MA, USA. <sup>9</sup>Department of Pediatrics, Harvard Medical School, Boston, MA, USA. <sup>145</sup>Present address: Centre for Population Genomics, Garvan Institute of Medical Research, and UNSW Sydney, Sydney, Australia. <sup>146</sup>Present address: Centre for Population Genomics, Murdoch Children's Research Institute, Melbourne, Australia. \*Lists of authors and their affiliations appear at the end of the paper. ✉e-mail: d.macarthur@garvan.org.au



**Fig. 1 | Curation of pLoF variants in haploinsufficient disease genes found in gnomAD reveals transcript errors as a major confounding error mode in variant annotation.** We identified and manually curated 401 pLoF variants in the gnomAD dataset in 61 haploinsufficient severe developmental delay genes and flagged any reason the pLoF may not be a true LoF variant. Top, the frequency of each error mode present in the 306 variants classified as unlikely

to be a true LoF. Transcript errors emerge as a major putative error mode in the annotation of these pLoF variants. Bottom, bee swarm plot shows the average pext score across GTEx tissues for each variant in the error categories. This shows that pext values are discriminately lower for variants that are annotated as possible transcript errors ( $P = 4.1 \times 10^{-38}$ , two-sided Wilcoxon test between transcript errors and other error modes).

Mendelian disease variants have been found on tissue-specific isoforms<sup>9,10</sup> and isoform expression levels in *TTN* have been used to show that pLoF variants found in healthy controls occur in exons that are absent from dominantly expressed isoforms, whereas those in patients with dilated cardiomyopathy occur on constitutive exons<sup>11</sup>, emphasizing the utility of exon expression information for variant interpretation.

### Isoform diversity and variant interpretation

We find that isoform diversity is a contributor to the paradoxical finding of disruptive variants in dosage-sensitive disease genes in ostensibly healthy individuals. In the gnomAD database, we identify 401 high-quality pLoF variants that pass both sequencing and annotation quality filters in 61 haploinsufficient disease genes in which heterozygous pLoF variants are established to cause severe developmental delay phenotypes with high penetrance (Methods). Given the severity of these phenotypes and their extremely low prevalence worldwide, ranging from 1 in 10,000 to less than 1 in a million, very few, if any true pLoF variants would be expected to be found in the gnomAD population. As such, most or all of these observed pLoF variants are likely to be sequencing or annotation errors<sup>12</sup>. Manual curation of these variants reveals common error modes that result in probable misannotation of pLoFs, with diversity of transcript structure, mediated by variants falling on low-confidence transcripts, emerging as a major consideration (Fig. 1, Supplementary Fig. 1, Supplementary Tables 1–3). However, no existing tools systematically incorporate information on transcript expression into variant interpretation.

### pext score summarizes isoform expression

The advent of large-scale transcriptome sequencing datasets, such as GTEx<sup>2</sup>, provides an opportunity to incorporate cross-tissue exon expression into variant interpretation. However, the current formats of these databases do not readily allow for unbiased estimation of exon expression. The GTEx web browser offers information on exon-level read pileup across tissues, but this approach is confounded by technical artefacts such as 3' bias<sup>13</sup> (preferential coverage of bases close to the 3' end of a transcript) (Supplementary Fig. 2a). Such systematic biases

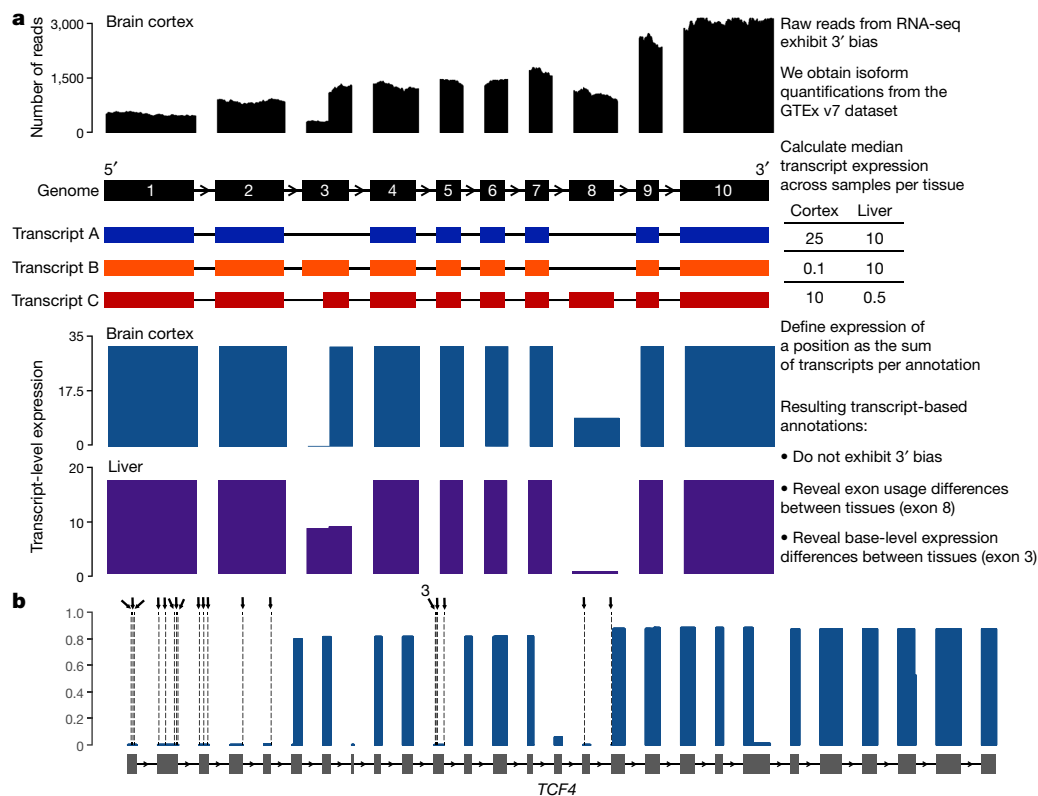
mean that simple exon-level coverage in a transcriptome dataset cannot be used as a reliable proxy for exon expression, especially in longer genes (Fig. 2a, Supplementary Fig. 2b).

Isoform quantification tools provide estimates of isoform expression levels that correct, albeit imperfectly<sup>13,14</sup>, for confounding by 3' bias as well as other technical artefacts such as isoform length, isoform GC content, and transcript sequence complexity<sup>15–17</sup>. Here, we use isoform-level quantifications from 11,706 tissue samples from the GTEx v7 dataset to derive an annotation-specific expression metric. For each tissue, we annotate each variant with the expression of every possible consequence across all transcripts, which can be used to summarize expression in any combination of tissues of interest. We first compute the median expression of a transcript across tissue samples, and define the expression of a given variant as the sum of the expression of all transcripts for which the variant has the same annotation (Fig. 2a, Supplementary Fig. 3a). By normalizing the expression of the annotation to the total gene expression, we define a metric (proportion expression across transcripts, or 'pext'), which can be interpreted as a measure of the proportion of the total transcriptional output from a gene that would be affected by the variant annotation in question (Supplementary Fig. 3b).

The pext metric allows for quick visualization of the expression of exons across a gene. In Fig. 2b, transcript-expression based annotation is shown for *TCF4*, a haploinsufficient gene in which heterozygous variants result in Pitt–Hopkins syndrome<sup>18</sup>, a highly penetrant disorder associated with severe developmental delay. This gene contains 20 unique high-quality pLoF mutations across 56 individuals in the gnomAD database. All 20 variants lie on exons with no evidence of expression across the GTEx dataset (Fig. 2b, Supplementary Fig. 4), which indicates that functional TCF4 protein can be made in the presence of these variants. This visualization is now available for all genes in the gnomAD browser (<https://gnomad.broadinstitute.org>), and can aid in the rapid identification of variants occurring on exons with little to no evidence of expression in GTEx.

### Functional validation of pext

To explore whether expression-based annotation marks functionally important regions, we compared the distribution of the pext metric in



**Fig. 2 | Summary of transcript-expression based annotation method.**

**a**, Overview of transcript-aware annotation. Most genes have many annotated isoforms, which can have varying expression patterns across tissues. Using the number of reads aligning to exonic regions in transcriptome datasets as a proxy for exon expression (top, black) has confounding effects, due to 3' bias. In this example, although exons 3 and 8 have markedly different expression levels in brain cortex, the number of reads aligning to the two exons is similar, and this masks the differences in exon usage. Transcript-aware annotation

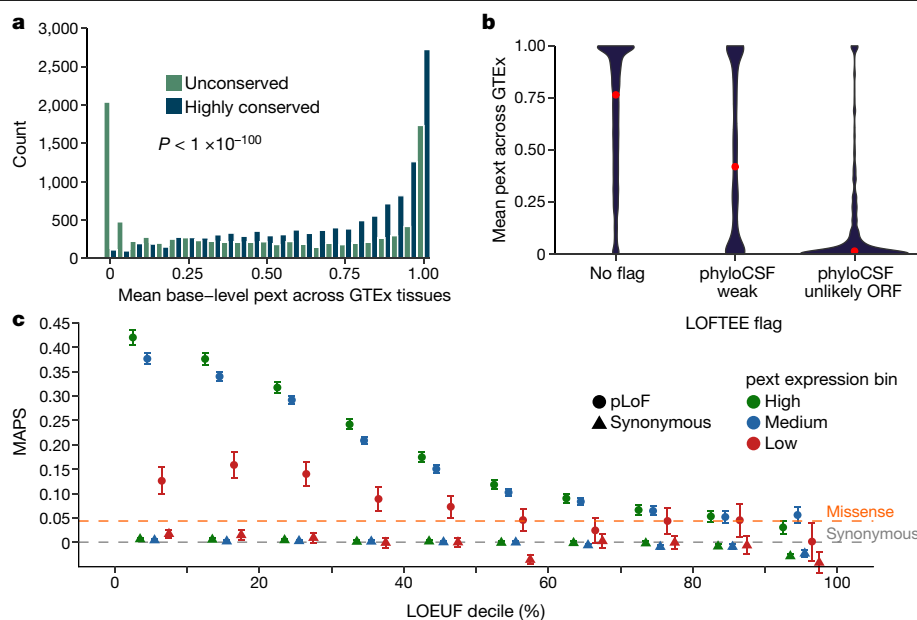
defines the expression of every variant as the sum of transcripts that have the same annotation. The resulting transcript-level expression plots do not exhibit 3' bias, and reveal differences in exon usage, such as those in exons 3 and 8, across tissues. **b**, Example of utility of transcript-expression based annotation. There are 20 high-quality pLoF variants in the haploinsufficient developmental delay gene *TCF4* in gnomAD, annotated as dashed lines and arrows. All 20 variants have no evidence of expression in the GTEx dataset, which suggests that functional TCF4 protein can be made in the presence of these variants.

evolutionarily conserved and unconserved regions using phyloCSF<sup>19</sup>. Exons with patterns of multi-species conservation consistent with coding regions have higher phyloCSF scores, and should exhibit detectable expression patterns, whereas regions with lower scores will be enriched for incorrect exon annotations, which are expected to have little evidence of expression in a population transcriptome dataset. As expected, we observe significantly lower expression for unconserved regions, and near-constitutive expression in highly conserved regions (Fig. 3a, Supplementary Fig. 5a). This difference remains statistically significant after correcting for exon length (logistic regression  $P < 1.0 \times 10^{-100}$ ), which can influence both phyloCSF scores and isoform quantifications, indicating that transcript expression-aware annotation marks functionally relevant exonic regions.

Although the metrics are associated, we find that pext provides orthogonal information to conservation for variant interpretation. For example, regions with low evidence of conservation but high expression (Fig. 3a) are enriched for genes in immune-related pathways (Methods), which are selected for diversity but represent true coding regions. In addition, the pext value is higher for pLoF variants annotated as high confidence by the loss-of-function transcript effect estimator (LOFTEE) package<sup>1</sup>, with no additional flags than those flagged as having found on unlikely open-reading frames or weakly conserved regions (Fig. 3b, Supplementary Fig. 5b). However, high-confidence LOFTEE variants with no flags can also have low pext values, which suggests that transcript-expression-aware annotation adds additional information to the currently available interpretation toolkit.

We undertook manual evaluation of 128 regions marked as unexpressed (mean pext < 0.1 in all tissues and in GTEx brain) in 61 haploinsufficient genes following the GENCODE manual annotation workflow<sup>20</sup> to evaluate the annotation quality in these coding sequence (CDS) regions. One-third of flagged regions were associated with low-quality models that have been removed or switched to non-coding biotypes in subsequent GENCODE releases (Supplementary Fig. 6), and 70% of the remaining regions correspond to models that satisfy only minimum criteria for inclusion in the gene set, corresponding to 'putative' annotations that lack markers for CDS functionality (Supplementary Table 4). Nonetheless, we find support for some highly conserved CDS regions, several of which show evidence of transcription in fetal tissues, underlining the importance of incorporating several isoform expression datasets for interpretation (Supplementary Fig. 6d).

Non-synonymous variants found on constitutively expressed regions would be expected to be more deleterious than those on regions with no evidence of expression. To test this, we defined expression bins based on the average pext value across GTEx tissues, in which an average pext value less than 0.1 was defined as low (or unexpressed), above 0.9 as high (or near-constitutive) and intermediate values as medium expression. We compared the mutability-adjusted proportion singleton (MAPS), a measure of negative selection on variant classes<sup>21</sup>, partitioned on the loss-of-function observed/expected upper-bound fraction (LOEUF) decile, a measure of constraint against pLoF variants in the gnomAD dataset<sup>1</sup> in each of these expression bins. MAPS scores differed substantially between pLoF variants found on low-expressed and high-expressed regions in genes intolerant to pLoF variation (Fig. 3c,



**Fig. 3 | Functional validation of transcript-expression based annotation.**  
**a**, We define highly conserved and unconserved regions as phyloCSF > 1,000 ( $n = 9,817$ ) and phyloCSF < -100 ( $n = 11,860$ ), respectively, and compare the expression status of these regions across GTEx. Regions with high phyloCSF scores are enriched for near-constitutive expression, whereas unconserved regions are enriched for little to no usage across GTEx. This difference is significant after correcting for gene length (logistic regression  $P < 1 \times 10^{-100}$ ). We note that unconserved regions with high levels of expression (pext > 0.9) are enriched for immune-related genes, which are selected for diversity and thus have low conservation, but represent true coding regions.  
**b**, Transcript-expression based annotation recapitulates, and adds information to, existing interpretation tools. High-confidence pLoF LOFTEE variants in gnomAD with no flags ( $n = 458,880$ ) are enriched for higher pext values, whereas high-confidence pLoF variants falling on low phyloCSF ( $n = 44,373$ ) or

unlikely open-reading frame regions ( $n = 2,437$ ) are enriched for low expression. However, high-confidence pLoF variants can also have a low pext score. Variants flagged falling on regions that are unlikely open-reading frame or have weak conservation are enriched for lower pext values. Red dots denote the median pext value across GTEx. **c**, Non-synonymous variants found on near-constitutive regions tend to be more deleterious. We compared the MAPS score for variants with low (<0.1), medium ( $0.1 \leq \text{pext} \leq 0.9$ ) and high (pext > 0.9) expression. Variants with near-constitutive expression have a higher MAPS score, which indicates higher deleteriousness than those with little to no evidence of expression. Points represent MAPS values and error bars denote the 95% confidence interval. Dashed grey and orange lines represent MAPS values for all gnomAD missense and synonymous variants, respectively. The number of variants evaluated per category and unadjusted proportion singleton values can be found in Supplementary Table 5a.

Supplementary Fig. 5c, Supplementary Table 5a, b). This information is complementary to existing variant prioritization tools such as PolyPhen-2<sup>22</sup> (Supplementary Fig. 5d, Supplementary Table 5c). This skew of non-synonymous variation in high-expressed regions suggests that variation arising in such exons tends to be more deleterious, whereas non-synonymous variants on regions with low expression are similar to missense variants in their inferred deleteriousness.

## Use of pext in variant interpretation

To evaluate the utility of transcript expression-based annotation in Mendelian variant interpretation, we assessed the number of variants that would be filtered based on a pext cut-off value of less than 0.1 (low expression) across GTEx tissues for three gene sets. First, we evaluated high-quality pLoF variants in the 61 manually curated haploinsufficient genes in gnomAD and ClinVar<sup>23</sup>. The low pext expression bin resulted in filtering of 22.8% of pLoF variants in haploinsufficient developmental delay genes in gnomAD, but only 3.8% of high-quality pathogenic variants in ClinVar ( $P = 4.7 \times 10^{-35}$ ) (Fig. 4a, Methods). We next compared pLoF variants in autosomal recessive disease genes found in a homozygous state in at least one individual in gnomAD and any pLoF variant in these genes in ClinVar and observed similar results: expression-based annotation filters 30.0% of variants in gnomAD while only filtering 3.2% of variants in ClinVar (Fig. 4b) ( $P = 3.5 \times 10^{-61}$ ).

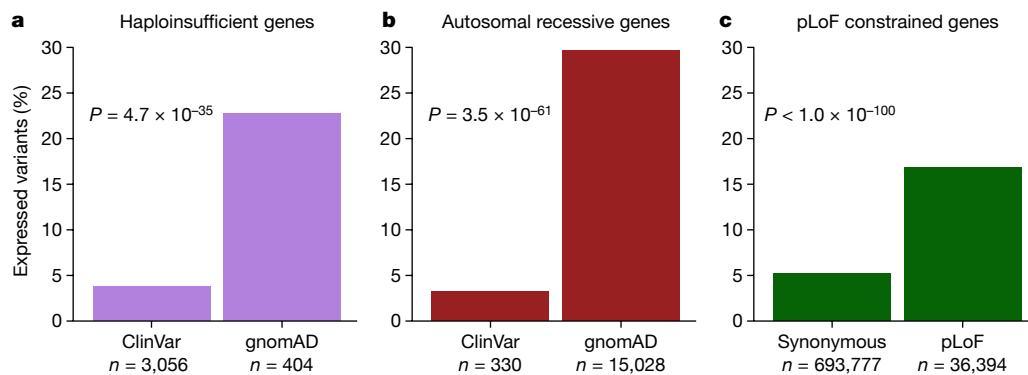
Finally, we evaluated gnomAD pLoF variants in genes that are constrained against pLoF variation<sup>1</sup> (LOEUF score < 0.35). Given that these genes are depleted for loss-of-function variation in the general population, we expect the observed pLoF variants in these genes to

be enriched for annotation errors. We compared the proportion filtered to synonymous variants in the same genes, which we expect to be randomly distributed. Our metric removes 16.8% of pLoF variants in constrained genes, but only 5.2% of synonymous variants (Fig. 4c) ( $P < 1.0 \times 10^{-100}$ ). In all cases, the vast majority of filtered variants were otherwise high-confidence with no LOFTEE annotation flags, which suggests again that pext provided additional information to existing variant prioritization tools in removing annotation errors (Supplementary Fig. 7).

## Use of pext in burden testing

To explore the benefits of this approach for rare variant analysis, we applied pext binning to burden testing of de novo variants in patients with developmental delay/intellectual disability (DD/ID) or autism spectrum disorder (ASD) using a set of 23,970 de novo variants collated from several studies including the Deciphering Developmental Disorders (DDD) project and the Autism Sequencing Consortium (ASC)<sup>24–29</sup>. We find that de novo pLoF variants in patients with DD/ID in low-expressed regions have similar effect sizes to those of synonymous variants (rate ratio of low-expressed pLoFs = 1.08,  $P = 0.90$ ), whereas pLoF variants in highly expressed regions have much larger effect sizes (rate ratio = 4.64,  $P = 3.74 \times 10^{-38}$ ) (Fig. 5a). This observation is consistent for de novo variants in autism (rate ratio for low-expressed pLoFs = 0.80,  $P = 0.47$ ; rate ratio for high-expressed pLoFs = 2.11,  $P = 8.2 \times 10^{-8}$ ) (Fig. 5b) and congenital heart disease with co-morbid neurodevelopmental delay (Supplementary Fig. 8a) as well as rare variants (allele count ≤ 10) identified in highly constrained genes in the large iPSYCH case-control study





**Fig. 4 | Transcript-expression based annotation aids Mendelian variant interpretation.** **a**, Comparison of the proportion of high-quality pLoF variants filtered in a curated list of 61 haploinsufficient developmental delay genes in gnomAD versus ClinVar with a cut-off value of average pext across GTEx  $\leq 0.1$  (low expression). Expression-based filtering results in removal of 22.8% of gnomAD pLoFs and 3.8% of confidently curated set of pLoFs in ClinVar. **b**, Expression-based annotation filters 30% of pLoF variants found in gnomAD

in a homozygous state in at least one individual, and 3.2% of any pLoF variants found in the same genes in ClinVar. **c**, We extended this filtering approach to pLoF and synonymous variants in gnomAD pLoF-intolerant genes (defined by LOEUF  $< 0.35$ ). This filters 16.8% of LoF and 5.2% of synonymous variants. The total number of high-quality variants considered in each group is shown. For all pLoFs only high-confidence LOFTEE variants were considered.  $P$  values were determined by two-sided Fisher's exact test for counts.

of Danish patients with autism spectrum disorder and attention-deficit/hyperactivity disorder (Supplementary Fig. 8b). Overall, we consistently observe low-expressed pLoFs to have effect sizes similar to those of synonymous variants, with pLoF variants in constitutive regions having larger effect sizes, which suggests that incorporating transcript expression-aware annotation in rare variant studies can boost power for gene discovery.

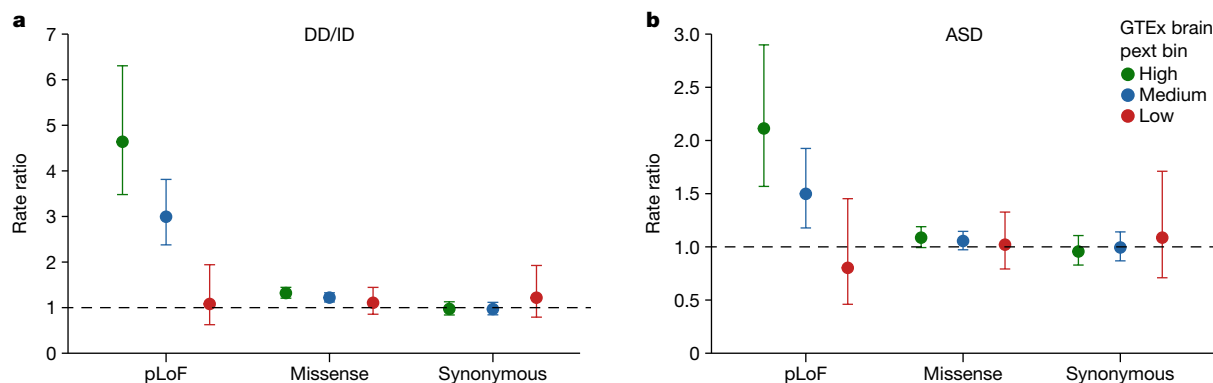
## Discussion

We have described the development and validation of a transcript expression-based annotation framework to integrate results from transcriptome sequencing experiments into clinical variant interpretation. Although our initial analysis uses GTEx, our method can be used with any isoform expression dataset to annotate any variant file rapidly in the scalable software framework Hail (<https://hail.is>). For example, annotation of more than 120,000 gnomAD individuals with GTEx takes under an hour using 60 cores, at a cost of about US\$5 on public cloud compute, which can be further scaled to larger datasets. In addition, the annotations we provide are flexible: although we have described the use of average transcript-level expression across many

tissues, alternative approaches such as using maximum expression across any tissue may prove useful depending on variant interpretation goals (Supplementary Figs. 9, 10).

We note that although this metric successfully discriminates between near-constitutive and low expression levels, which are useful for prioritizing and filtering variants, respectively, regions with intermediate expression levels are more challenging to interpret. However, we hypothesize directed analyses of intermediate expression levels may help to determine the role of alternative splicing in phenotypic diversity<sup>30,31</sup>. In addition, although we have binned average pext scores across GTEx tissues into low, medium and high expression, different genes will probably have varying optimal tissues and thresholds for variant interpretation. Regions tagged as low expression are often corroborated by expert opinion of CDS curation, but domain knowledge of a gene will outperform this summary metric.

An important caveat in our approach is the imprecision of isoform quantification methods using short-read transcriptome data. However, we note that repeating key analyses in the manuscript with a different isoform quantification tool showed consistent results (Methods, Supplementary Fig. 11, Supplementary Table 6), suggesting robustness to the precise pipeline used. The utility of this framework will increase



**Fig. 5 | Application of transcript-expression based annotation to de novo variant analyses in ASD and DD/ID.** **a**, **b**, Transcript-expression-based analyses in patients with DD/ID (**a**) or ASD (**b**). We find that de novo pLoF variants found on near-constitutively expressed regions in GTEx brain tissues have larger effect sizes than de novo LoF variants in weakly expressed regions in both disorders. Notably, de novo pLoF variants found on regions with little evidence for expression are as equally distributed in cases versus controls as

de novo synonymous variants, which suggests that such variants can be removed from analyses of gene burden testing to boost discovery power. The high pext expression bin contains 46.1%, 42.3% and 11.4%, and the low-expression bin contains 4.0%, 6.0% and 11.4% of 1,249, 752 and 166 de novo pLoF variants found in patients with DD/ID, ASD and controls, respectively. Points represent rate ratio estimate and error bars represent 95% confidence interval from the Poisson exact test.

as our ability to quantify isoform expression across tissues improves, including refinement of methods and gene models, as well as availability of long-read RNA-sequencing data from human tissues. In addition, the improvement of single-cell RNA-seq technologies and the generation of data across human tissues will provide insight into cell type-specific exon usage for incorporation into variant interpretation<sup>32</sup>.

The code used to generate pext is available as open source software ([https://github.com/macarthur-lab/tx\\_annotation](https://github.com/macarthur-lab/tx_annotation)). In addition, we provide a precomputed file of the transcript expression value for every possible single nucleotide variant in the human genome. This metric has already proven useful in variant curation for the identification of drug targets<sup>33</sup> and for filtering variants for the identification of human knockouts<sup>1</sup>. Overall, our metric can be incorporated into variant interpretation in Mendelian disease pipelines, analyses of rare variant burden, and the prioritization of variants for recall-by-genotype studies.

## Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-020-2329-2>.

- Karczewski, K. J. et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* <https://doi.org/10.1038/s41586-020-2308-7> (2020).
- GTEx Consortium. Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).
- MacArthur, D. G. et al. Guidelines for investigating causality of sequence variants in human disease. *Nature* **508**, 469–476 (2014).
- Goldstein, D. B. et al. Sequencing studies in human genetics: design and interpretation. *Nat. Rev. Genet.* **14**, 460–470 (2013).
- Dick, I. E., Joshi-Mukherjee, R., Yang, W. & Yue, D. T. Arrhythmogenesis in Timothy Syndrome is associated with defects in Ca<sup>2+</sup>-dependent inactivation. *Nat. Commun.* **7**, 10370 (2016).
- Splawski, I. et al. Ca<sub>v</sub>1.2 calcium channel dysfunction causes a multisystem disorder including arrhythmia and autism. *Cell* **119**, 19–31 (2004).
- Liao, P. & Soong, T. W. CaV1.2 channelopathies: from arrhythmias to autism, bipolar disorder, and immunodeficiency. *Pharmacol. Ther.* **107**, 353–359 (2010).
- Splawski, I. et al. Severe arrhythmia disorder caused by cardiac L-type calcium channel mutations. *Proc. Natl Acad. Sci. USA* **102**, 8089–8096 (2005).
- Nousbeck, J. et al. A mutation in a skin-specific isoform of SMARCA1 causes autosomal-dominant adermatoglyphia. *Am. J. Hum. Genet.* **89**, 302–307 (2011).
- Guven, A. & Tolun, A. TBC1D24 truncating mutation resulting in severe neurodegeneration. *J. Med. Genet.* **50**, 199–202 (2013).
- Roberts, A. M. et al. Integrated allelic, transcriptional, and phenomic dissection of the cardiac effects of titin truncations in health and disease. *Sci. Transl. Med.* **7**, 270ra276 (2015).
- MacArthur, D. G. et al. A systematic survey of loss-of-function variants in human protein-coding genes. *Science* **335**, 823–828 (2012).
- Zheng, W., Chung, L. M. & Zhao, H. J. B. B. Bias detection and correction in RNA-sequencing data. *BMC Bioinformatics* **12**, 290 (2011).
- Teng, M. et al. A benchmark for RNA-seq quantification pipelines. *Genome Biol.* **17**, 74 (2016).
- Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323 (2011).
- Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **34**, 525–527 (2016).
- Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* **14**, 417–419 (2017).
- Sweatt, J. D. Pitt-Hopkins Syndrome: intellectual disability due to loss of TCF4-regulated gene transcription. *Mol. Med.* **45**, e21 (2013).
- Lin, M. F., Jungreis, I. & Kellis, M. PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics* **27**, i275–i282 (2011).
- Frankish, A. et al. GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res.* **47** (D1), D766–D773 (2019).
- Lek, M. et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).
- Adzhubei, I., Jordan, D. M. & Sunyaev, S. R. Predicting functional effect of human missense mutations using PolyPhen-2. *Curr. Prot. Human Genet.* **Chapter 7**, Unit7.20–Unit7.20 (2013).
- Landrum, M. J. et al. ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.* **46** (D1), D1062–D1067 (2018).
- Deciphering Developmental Disorders Study. Prevalence and architecture of *de novo* mutations in developmental disorders. *Nature* **542**, 433–438 (2017).
- Satterstrom, F. K. et al. Large-scale exome sequencing study implicates both developmental and functional changes in the neurobiology of autism. *Cell* **180**, 568–584. e23 (2020).

- Lelieveld, S. H. et al. Meta-analysis of 2,104 trios provides support for 10 new genes for intellectual disability. *Nat. Neurosci.* **19**, 1194–1196 (2016).
- Hamdan, F. F. et al. De novo mutations in moderate or severe intellectual disability. *PLoS Genet.* **10**, e1004772–e1004772 (2014).
- de Ligt, J. et al. Diagnostic exome sequencing in persons with severe intellectual disability. *N. Engl. J. Med.* **367**, 1921–1929 (2012).
- Rauch, A. et al. Range of genetic mutations associated with severe non-syndromic sporadic intellectual disability: an exome sequencing study. *Lancet* **380**, 1674–1682 (2012).
- Tress, M. L., Abascal, F. & Valencia, A. Alternative splicing may not be the key to proteome complexity. *Trends Biochem. Sci.* **42**, 98–110 (2017).
- Blencowe, B. J. The relationship between alternative splicing and proteomic complexity. *Trends Biochem. Sci.* **42**, 407–408 (2017).
- Regev, A. et al. The Human Cell Atlas. *eLife*. **6**, e27041 (2017).
- Minikel, E. V. et al. Evaluating drug targets through human loss-of-function genetic variation. *Nature* <https://doi.org/10.1038/s41586-020-2267-z> (2019).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2020

## Genome Aggregation Database Production Team

Jessica Alföldi<sup>1,2</sup>, Irina M. Armean<sup>1,2,6</sup>, Eric Banks<sup>10</sup>, Louis Bergelson<sup>10</sup>, Kristian Cibulskis<sup>10</sup>, Ryan L. Collins<sup>1,4,11</sup>, Kristen M. Connolly<sup>12</sup>, Miguel Covarrubias<sup>10</sup>, Beryl B. Cummings<sup>1,2,3</sup>, Mark J. Daly<sup>1,2</sup>, Stacey Donnelly<sup>1</sup>, Yossi Farjoun<sup>10</sup>, Steven Ferreira<sup>13</sup>, Laurent Franciolli<sup>1,2</sup>, Stacey Gabriel<sup>13</sup>, Laura D. Gauthier<sup>10</sup>, Jeff Gentry<sup>10</sup>, Namrata Gupta<sup>1,13</sup>, Thibault Jeandet<sup>10</sup>, Diane Kaplan<sup>10</sup>, Konrad J. Karczewski<sup>1,2</sup>, Kristen M. Laricchia<sup>1,2</sup>, Christopher Llanwarne<sup>10</sup>, Eric V. Minikel<sup>1,2,7</sup>, Ruchi Munshi<sup>10</sup>, Benjamin M. Neale<sup>1,2</sup>, Sam Novod<sup>10</sup>, Anne H. O'Donnell-Luria<sup>1,8,9</sup>, Nikelle Petrillo<sup>10</sup>, Timothy Poterba<sup>1,2,7</sup>, David Roazen<sup>10</sup>, Valentin Ruano-Rubio<sup>10</sup>, Andrea Saltzman<sup>1</sup>, Kaitlin E. Samocha<sup>1,4</sup>, Molly Schleicher<sup>1</sup>, Cotton Seed<sup>10</sup>, Matthew Solomonson<sup>1,2</sup>, Jose Soto<sup>10</sup>, Grace Tiao<sup>1,2</sup>, Kathleen Tibbetts<sup>10</sup>, Charlotte Tolonen<sup>10</sup>, Christopher Vittal<sup>1,2,7</sup>, Gordon Wade<sup>10</sup>, Arcturus Wang<sup>1,2,7</sup>, Qingbo Wang<sup>1,2,4</sup>, James S. Ware<sup>1,15,16</sup>, Nicholas A. Watts<sup>1,2</sup>, Ben Weisburd<sup>10</sup> & Nicola Whiffin<sup>1,15,16</sup>

<sup>10</sup>Data Sciences Platform, Broad Institute of MIT and Harvard, Cambridge, MA, USA. <sup>11</sup>Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA, USA. <sup>12</sup>Genomics Platform, Broad Institute of MIT and Harvard, Cambridge, MA, USA. <sup>13</sup>Broad Genomics, Broad Institute of MIT and Harvard, Cambridge, MA, USA. <sup>14</sup>Wellcome Sanger Institute, Hinxton, Cambridgeshire, UK. <sup>15</sup>National Heart & Lung Institute and MRC London Institute of Medical Sciences, Imperial College London, London, UK. <sup>16</sup>Cardiovascular Research Centre, Royal Brompton & Harefield Hospitals NHS Trust, London, UK.

## Genome Aggregation Database Consortium

Carlos A. Aguilar Salinas<sup>17</sup>, Tariq Ahmad<sup>18</sup>, Christine M. Albert<sup>19,20</sup>, Diego Ardisson<sup>21</sup>, Gil Atzmon<sup>22,23,24</sup>, John Barnard<sup>25</sup>, Laurent Beaugier<sup>26</sup>, Emelia J. Benjamin<sup>27,28,29</sup>, Michael Boehnke<sup>30</sup>, Lori L. Bonnycastle<sup>31</sup>, Erwin P. Bottinger<sup>32</sup>, Donald W. Bowden<sup>33,34,35</sup>, Matthew J. Bown<sup>36,37</sup>, John C. Chambers<sup>38,39,40</sup>, Juliana C. Chan<sup>41</sup>, Daniel Chasman<sup>19</sup>, Judy Cho<sup>32</sup>, Mina K. Chung<sup>25</sup>, Bruce Cohen<sup>42,43</sup>, Adolfo Correa<sup>44</sup>, Dana Dabealea<sup>45</sup>, Mark J. Daly<sup>1,2</sup>, Dawood Darbar<sup>46</sup>, Ravindranath Duggirala<sup>47</sup>, José Dupuis<sup>27,48</sup>, Patrick T. Ellinor<sup>49</sup>, Roberto Elosua<sup>50,51,52</sup>, Jeanette Erdmann<sup>53,54,55</sup>, Tõnu Esko<sup>56</sup>, Martti Färkkilä<sup>57</sup>, Jose Florez<sup>1</sup>, Andre Franke<sup>58</sup>, Gad Getz<sup>59</sup>, Benjamin Glaser<sup>60</sup>, Stephen J. Glatt<sup>61</sup>, David Goldstein<sup>62,63</sup>, Clicerio Gonzalez<sup>64</sup>, Leif Groop<sup>65</sup>, Christopher Haiman<sup>66</sup>, Craig Hanis<sup>67</sup>, Matthew Harms<sup>68,69</sup>, Mikko Hiltunen<sup>70</sup>, Matti M. Hoti<sup>71</sup>, Christina M. Hultman<sup>72</sup>, Mikko Kallela<sup>73</sup>, Jaakko Kaprio<sup>74,75</sup>, Sekar Kathiresan<sup>1,17,76</sup>, Bong-Jo Kim<sup>77</sup>, Young Jin Kim<sup>77</sup>, George Kirov<sup>78</sup>, Jaspal Kooner<sup>38,40,79</sup>, Seppo Koskeno<sup>80</sup>, Harlan M. Krumholz<sup>81</sup>, Subra Kugathasan<sup>82</sup>, Soo Heon Kwak<sup>83</sup>, Markku Laakso<sup>84,85</sup>, Terho Lehtimäki<sup>86</sup>, Ruth J. F. Loos<sup>32,87</sup>, Steven A. Lubitz<sup>1,88</sup>, Ronald C. W. Ma<sup>89,90,91</sup>, Daniel G. MacArthur<sup>1,2</sup>, Jaume Marrugat<sup>91,92</sup>, Kari M. Mattila<sup>86</sup>, Steven McCarroll<sup>73,93</sup>, Mark I. McCarthy<sup>94,95,96</sup>, Dermot McGovern<sup>97</sup>, Ruth McPherson<sup>98</sup>, James B. Meigs<sup>1,76,99</sup>, Olle Melander<sup>100</sup>, Andres Metspalu<sup>56</sup>, Benjamin M. Neale<sup>1,2</sup>, Peter M. Nilsson<sup>101</sup>, Michael C. O'Donovan<sup>78</sup>, Dost Ongur<sup>42,76</sup>, Lorena Orozco<sup>102</sup>, Michael J. Owen<sup>78</sup>, Colin N. A. Palmer<sup>103</sup>, Aarno Palotie<sup>1,2,74</sup>, Kyong Soo Park<sup>83,104</sup>, Carlos Pato<sup>105</sup>, Ann E. Pulver<sup>106</sup>, Nazneen Rahman<sup>107</sup>, Anne M. Remes<sup>108</sup>, John D. Rioux<sup>109,110</sup>, Samuli Ripatti<sup>1,74,111</sup>, Dan M. Roden<sup>112,113</sup>, Danish Saleheen<sup>114,115,116</sup>, Veikko Salomaa<sup>117</sup>, Nilesh J. Samani<sup>36,37</sup>, Jeremiah Scharf<sup>1,71</sup>, Heribert Schunkert<sup>118,119</sup>, Moore B. Shoemaker<sup>120</sup>, Pamela Sklar<sup>121,122,123,127</sup>, Hilkkka Soininen<sup>124</sup>,

**Harry Sokol**<sup>26</sup>, **Tim Spector**<sup>125</sup>, **Patrick F. Sullivan**<sup>72,126</sup>, **Jaana Suvisaari**<sup>117</sup>, **E. Shyong Tai**<sup>127,128,129</sup>, **Yik Ying Teo**<sup>127,130,131</sup>, **Tuomi Tiinamäla**<sup>74,132,133</sup>, **Ming Tsuang**<sup>134,135</sup>, **Dan Turner**<sup>136</sup>, **Teresa Tusie-Luna**<sup>137,138</sup>, **Erkki Vartiainen**<sup>139</sup>, **James S. Ware**<sup>135,136</sup>, **Hugh Watkins**<sup>140</sup>, **Rinse K. Weersma**<sup>141</sup>, **Maija Wessman**<sup>74,132</sup>, **James G. Wilson**<sup>142</sup> & **Ramnik J. Xavier**<sup>143,144</sup>

<sup>17</sup>Unidad de Investigacion de Enfermedades Metabolicas, Instituto Nacional de Ciencias Medicas y Nutricion, Mexico City, Mexico. <sup>18</sup>Peninsula College of Medicine and Dentistry, Exeter, UK. <sup>19</sup>Division of Preventive Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA. <sup>20</sup>Division of Cardiovascular Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA. <sup>21</sup>Department of Cardiology, University Hospital, Parma, Italy. <sup>22</sup>Department of Biology, Faculty of Natural Sciences, University of Haifa, Haifa, Israel. <sup>23</sup>Department of Medicine, Albert Einstein College of Medicine, Bronx, NY, USA. <sup>24</sup>Department of Genetics, Albert Einstein College of Medicine, Bronx, NY, USA. <sup>25</sup>Department of Quantitative Health Sciences, Lerner Research Institute, Cleveland Clinic, Cleveland, OH, USA. <sup>26</sup>Sorbonne Université, APHP, Gastroenterology Department, Saint Antoine Hospital, Paris, France. <sup>27</sup>Framingham Heart Study, National Heart, Lung, & Blood Institute and Boston University, Framingham, MA, USA. <sup>28</sup>Department of Medicine, Boston University School of Medicine, Boston, MA, USA. <sup>29</sup>Department of Epidemiology, Boston University School of Public Health, Boston, MA, USA. <sup>30</sup>Department of Biostatistics, Center for Statistical Genetics, University of Michigan, Ann Arbor, MI, USA. <sup>31</sup>National Human Genome Research Institute, National Institutes of Health, Bethesda, MD, USA. <sup>32</sup>The Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, USA. <sup>33</sup>Department of Biochemistry, Wake Forest School of Medicine, Winston-Salem, NC, USA. <sup>34</sup>Center for Genomics and Personalized Medicine Research, Wake Forest School of Medicine, Winston-Salem, NC, USA. <sup>35</sup>Center for Diabetes Research, Wake Forest School of Medicine, Winston-Salem, NC, USA. <sup>36</sup>Department of Cardiovascular Sciences and NIHR Leicester Biomedical Research Centre, University of Leicester, Leicester, UK. <sup>37</sup>NIHR Leicester Biomedical Research Centre, Glenfield Hospital, Leicester, UK. <sup>38</sup>Department of Epidemiology and Biostatistics, Imperial College London, London, UK. <sup>39</sup>Department of Cardiology, Ealing Hospital NHS Trust, Southall, UK. <sup>40</sup>Imperial College Healthcare NHS Trust, Imperial College London, London, UK. <sup>41</sup>Department of Medicine and Therapeutics, The Chinese University of Hong Kong, Hong Kong, Hong Kong. <sup>42</sup>Program for Neuropsychiatric Research, McLean Hospital, Belmont, MA, USA. <sup>43</sup>Department of Psychiatry, Harvard Medical School, Boston, MA, USA. <sup>44</sup>Department of Medicine, University of Mississippi Medical Center, Jackson, MI, USA. <sup>45</sup>Department of Epidemiology, Colorado School of Public Health, Aurora, CP, USA. <sup>46</sup>Department of Medicine and Pharmacology, University of Illinois at Chicago, Chicago, IL, USA. <sup>47</sup>Department of Genetics, Texas Biomedical Research Institute, San Antonio, TX, USA. <sup>48</sup>Department of Biostatistics, Boston University School of Public Health, Boston, MA, USA. <sup>49</sup>Cardiac Arrhythmia Service and Cardiovascular Research Center, Massachusetts General Hospital, Boston, MA, USA. <sup>50</sup>Cardiovascular Epidemiology and Genetics, Hospital del Mar Medical Research Institute (IMIM), Barcelona, Spain. <sup>51</sup>Centro de Investigación Biomédica en Red Enfermedades Cardiovasculares (CIBERCV), Barcelona, Spain. <sup>52</sup>Department of Medicine, Medical School, University of Vic-Central University of Catalonia, Vic, Spain. <sup>53</sup>Institute for Cardiogenetics, University of Lübeck, Lübeck, Germany. <sup>54</sup>DZHK (German Research Centre for Cardiovascular Research), partner site Hamburg/Lübeck/Kiel, Lübeck, Germany. <sup>55</sup>University Heart Center Lübeck, Lübeck, Germany. <sup>56</sup>Estonian Genome Center, Institute of Genomics, University of Tartu, Estonia. <sup>57</sup>Clinic of Gastroenterology, Helsinki University and Helsinki University Hospital, Helsinki, Finland. <sup>58</sup>Institute of Clinical Molecular Biology (IKMB), Christian-Albrechts-University of Kiel, Kiel, Germany. <sup>59</sup>Cancer Genome Computational Analysis Group, Broad Institute of MIT and Harvard, Cambridge, MA, USA. <sup>60</sup>Endocrinology and Metabolism Department, Hadassah-Hebrew University Medical Center, Jerusalem, Israel. <sup>61</sup>Department of Psychiatry and Behavioral Sciences, SUNY Upstate Medical University, Syracuse, NY, USA. <sup>62</sup>Institute for Genomic Medicine, Columbia University Medical Center, Hammer Health Sciences, New York, NY, USA. <sup>63</sup>Department of Genetics & Development, Columbia University Medical Center, Hammer Health Sciences, New York, NY, USA. <sup>64</sup>Centro de Investigacion en Salud Poblacional, Instituto Nacional de Salud Publica, Cuernavaca, Mexico. <sup>65</sup>Genomics, Diabetes and Endocrinology, Lund University, Lund, Sweden. <sup>66</sup>Lund University Diabetes Centre, Malmö, Sweden. <sup>67</sup>Human Genetics Center, University of Texas Health Science Center at Houston, Houston, TX, USA. <sup>68</sup>Department of Neurology, Columbia University, New York, NY, USA. <sup>69</sup>Institute of Genomic Medicine, Columbia University, New York, NY, USA. <sup>70</sup>Institute of Biomedicine, University of Eastern Finland, Kuopio, Finland. <sup>71</sup>Department of Psychiatry, Helsinki University Central Hospital, Lapinlahdentie, Helsinki, Finland. <sup>72</sup>Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden. <sup>73</sup>Department of Neurology, Helsinki University Central Hospital, Helsinki, Finland. <sup>74</sup>Institute for Molecular Medicine FIMM, University of Helsinki, Helsinki, Finland. <sup>75</sup>Department of Public Health, University of Helsinki, Helsinki, Finland. <sup>76</sup>Department of Medicine, Harvard Medical School, Boston, MA, USA. <sup>77</sup>Center for Genome Science, Korea National Institute of Health, Chungcheongbuk-do, South Korea. <sup>78</sup>MRC Centre for Neuropsychiatric Genetics & Genomics, Cardiff University School of Medicine, Cardiff, UK. <sup>79</sup>National Heart and Lung Institute and MRC London Institute of Medical Sciences, Imperial College London, London, UK. <sup>80</sup>Department of Health, THL-National Institute for Health and

Welfare, Helsinki, Finland. <sup>81</sup>Section of Cardiovascular Medicine, Department of Internal Medicine, Yale School of Medicine, New Haven, CT, USA. <sup>82</sup>Division of Pediatric Gastroenterology, Emory University School of Medicine, Atlanta, Georgia, USA. <sup>83</sup>Department of Internal Medicine, Seoul National University Hospital, Seoul, South Korea. <sup>84</sup>The University of Eastern Finland, Institute of Clinical Medicine, Kuopio, Finland. <sup>85</sup>Kuopio University Hospital, Kuopio, Finland. <sup>86</sup>Department of Clinical Chemistry, Fimlab Laboratories and Finnish Cardiovascular Research Center-Tampere, Faculty of Medicine and Health Technology, Tampere University, Tampere, Finland. <sup>87</sup>The Mindich Child Health and Development Institute, Icahn School of Medicine at Mount Sinai, New York, NY, USA. <sup>88</sup>Cardiac Arrhythmia Service, Massachusetts General Hospital, Boston, MA, USA. <sup>89</sup>Department of Medicine and Therapeutics, The Chinese University of Hong Kong, Hong Kong, China. <sup>90</sup>Li Ka Shing Institute of Health Sciences, The Chinese University of Hong Kong, Hong Kong, China. <sup>91</sup>Hong Kong Institute of Diabetes and Obesity, The Chinese University of Hong Kong, Hong Kong, China. <sup>92</sup>Cardiovascular Research REGICOR Group, Hospital del Mar Medical Research Institute (IMIM), Barcelona, Spain. <sup>93</sup>Department of Genetics, Harvard Medical School, Boston, MA, USA. <sup>94</sup>Oxford Centre for Diabetes, Endocrinology and Metabolism, University of Oxford, Churchill Hospital, Oxford, UK. <sup>95</sup>Wellcome Centre for Human Genetics, University of Oxford, Oxford, UK. <sup>96</sup>Oxford NIHR Biomedical Research Centre, Oxford University Hospitals NHS Foundation Trust, John Radcliffe Hospital, Oxford, UK. <sup>97</sup>F Widjaja Foundation Inflammatory Bowel and Immunobiology Research Institute, Cedars-Sinai Medical Center, Los Angeles, CA, USA. <sup>98</sup>Atherogenomics Laboratory, University of Ottawa Heart Institute, Ottawa, Canada. <sup>99</sup>Division of General Internal Medicine, Massachusetts General Hospital, Boston, MA, USA. <sup>100</sup>Department of Clinical Sciences, University Hospital Malmö Clinical Research Center, Lund University, Malmö, Sweden. <sup>101</sup>Lund University, Dept. Clinical Sciences, Skane University Hospital, Malmö, Sweden. <sup>102</sup>Instituto Nacional de Medicina Genómica (INMEGEN), Mexico City, Mexico. <sup>103</sup>Medical Research Institute, Ninewells Hospital and Medical School, University of Dundee, Dundee, UK. <sup>104</sup>Department of Molecular Medicine and Biopharmaceutical Sciences, Graduate School of Convergence Science and Technology, Seoul National University, Seoul, South Korea. <sup>105</sup>Department of Psychiatry, Keck School of Medicine at the University of Southern California, Los Angeles, CA, USA. <sup>106</sup>Department of Psychiatry and Behavioral Sciences, Johns Hopkins University School of Medicine, Baltimore, MD, USA. <sup>107</sup>Division of Genetics and Epidemiology, Institute of Cancer Research, London, UK. <sup>108</sup>Research Unit of Clinical Neuroscience, University of Oulu, Oulu, Finland. <sup>109</sup>Research Center, Montreal Heart Institute, Montreal, Quebec, Canada. <sup>110</sup>Department of Medicine, Faculty of Medicine, Université de Montréal, Montreal, Quebec, Canada. <sup>111</sup>Department of Public Health, Faculty of Medicine, University of Helsinki, Helsinki, Finland. <sup>112</sup>Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN, USA. <sup>113</sup>Department of Medicine, Vanderbilt University Medical Center, Nashville, TN, USA. <sup>114</sup>Department of Biostatistics and Epidemiology, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA, USA. <sup>115</sup>Department of Medicine, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA, USA. <sup>116</sup>Center for Non-Communicable Diseases, Karachi, Pakistan. <sup>117</sup>National Institute for Health and Welfare, Helsinki, Finland. <sup>118</sup>Deutsches Herzzentrum München, Munich, Germany. <sup>119</sup>Technische Universität München, Munich, Germany. <sup>120</sup>Division of Cardiovascular Medicine, Nashville VA Medical Center and Vanderbilt University, School of Medicine, Nashville, TN, USA. <sup>121</sup>Department of Psychiatry, Icahn School of Medicine at Mount Sinai, New York, NY, USA. <sup>122</sup>Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY, USA. <sup>123</sup>Institute for Genomics and Multiscale Biology, Icahn School of Medicine at Mount Sinai, New York, NY, USA. <sup>124</sup>Institute of Clinical Medicine Neurology, University of Eastern Finland, Kuopio, Finland. <sup>125</sup>Department of Twin Research and Genetic Epidemiology, King's College London, London, UK. <sup>126</sup>Departments of Genetics and Psychiatry, University of North Carolina, Chapel Hill, NC, USA. <sup>127</sup>Saw Swee Hock School of Public Health, National University of Singapore, National University Health System, Singapore, Singapore. <sup>128</sup>Department of Medicine, Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore. <sup>129</sup>Duke-NUS Graduate Medical School, Singapore, Singapore. <sup>130</sup>Life Sciences Institute, National University of Singapore, Singapore, Singapore. <sup>131</sup>Department of Statistics and Applied Probability, National University of Singapore, Singapore, Singapore. <sup>132</sup>Folkhälsan Institute of Genetics, Folkhälsan Research Center, Helsinki, Finland. <sup>133</sup>HUCH Abdominal Center, Helsinki University Hospital, Helsinki, Finland. <sup>134</sup>Center for Behavioral Genomics, Department of Psychiatry, University of California, San Diego, CA, USA. <sup>135</sup>Institute of Genomic Medicine, University of California, San Diego, CA, USA. <sup>136</sup>Juliet Keidan Institute of Pediatric Gastroenterology, Shaare Zedek Medical Center, The Hebrew University of Jerusalem, Jerusalem, Israel. <sup>137</sup>Instituto de Investigaciones Biomédicas UNAM, Mexico City, Mexico. <sup>138</sup>Instituto Nacional de Ciencias Médicas y Nutrición Salvador Zubirán, Mexico City, Mexico. <sup>139</sup>Department of Public Health Solutions, National Institute for Health and Welfare, Helsinki, Finland. <sup>140</sup>Radcliffe Department of Medicine, University of Oxford, Oxford, UK. <sup>141</sup>Department of Gastroenterology and Hepatology, University of Groningen and University Medical Center Groningen, Groningen, The Netherlands. <sup>142</sup>Department of Physiology and Biophysics, University of Mississippi Medical Center, Jackson, MS, USA. <sup>143</sup>Program in Infectious Disease and Microbiome, Broad Institute of MIT and Harvard, Cambridge, MA, USA. <sup>144</sup>Center for Computational and Integrative Biology, Massachusetts General Hospital, Boston, MA, USA. <sup>147</sup>Deceased: Pamela Sklar.



## Methods

### Data reporting

No statistical methods were used to predetermine sample size. The experiments were not randomized, and investigators were not blinded to allocation during experiments and outcome assessment.

### Curation of pLoF variants in haploinsufficient developmental disease genes

To identify haploinsufficient developmental delay genes, we selected genes curated by the ClinGen Dosage Sensitivity Working Group<sup>34</sup>. 58 of the 61 genes had a score of 3 with sufficient evidence for pathogenicity, whereas two genes (*CHMP1*, *CTCF*) had a score of 2 (some evidence) and one gene (*RERE*) was not yet scored. The penetrance of pathogenic variants in each gene was reviewed in the literature, and only genes with more than 75% reported penetrance were included. These conditions are those too severe to expect to see an individual in gnomAD (probably unable to consent for a study without guardianship). The 61 genes include 50 autosomal genes of high severity and high penetrance and 11 genes on chromosome X in which the phenotype is expected to be severe or lethal in males and moderate to severe in females. The resulting gene list is available at [gs://gnomad-public/papers/2019-tx-annotation/data/gene\\_lists/HI\\_genes\\_100417.tsv](https://gnomad-public/papers/2019-tx-annotation/data/gene_lists/HI_genes_100417.tsv).

We extracted pLoF variants, defined as essential splice acceptor, essential splice donor, stop-gained, and frameshift variants, identified in the 61 haploinsufficient disease genes from the gnomAD v2.1.1 exome and genome sites tables, and considered only those pLoF variants that passed random forest filtering in the gnomAD dataset, and were annotated as high confidence by LOFTEE v1.0. Of 61 genes, 55 had at least one high-quality pLoF available in gnomAD. We performed manual curation of 401 pLoF variants using a web-based curation portal to identify any reason a pLoF may have been a variant calling or annotation error, and categorized the likelihood of each variant being a true LoF.

Evidence for classifying an LoF variant as artefactual was categorized into the following groups: mapping error, strand bias, reference error, genotyping error, homopolymer sequence, in-frame multi-nucleotide variant or frame-restoring indel, essential splice site rescue, minority of transcripts, weak exon conservation, last exon, and other annotation error. All possible reasons also to reject a LoF consequence were flagged, even when a single criterion would categorize the variant as not LoF. Variants were then categorized as LoF, probable LoF, probably not LoF, and not LoF based on criteria outlined in Supplementary Table 2. Supplementary Fig. 1a shows the distribution of the LoF verdicts for the 401 pLoF variants.

Technical errors comprised genotyping errors, strand biases, reference errors, and repetitive regions that could be detected by visual inspection of reads in the Integrative Genomics Viewer<sup>35</sup> (IGV) and from the UCSC genome browser<sup>36</sup>. Genotyping errors comprised skewed allele balances (conservative cutoff of  $\leq 35\%$ ), low complexity sequences, GC-rich regions, homopolymer tracts ( $\geq 6$  base pairs or  $\geq 6$  trinucleotide repeats) and low quality metrics (genotype quality  $< 20$ ). Strand bias was flagged when a variant was skewed preferentially on the forward or reverse strand, or when the majority ( $> 90\%$ ) of a given strand covered a region; this was often observed around intron–exon boundaries. Strand biases despite balanced coverage of the forward and reverse strands were weighted towards probably not LoF, whereas a strand bias due to skewed strand coverage was weighted alongside other genotyping errors. Reference errors were uncommon, but identified by a small deletion in a given exon, posing as a  $< 5$ -base-pair intron. Most genotyping errors and strand biases in isolation were not deemed critical in deciding whether a variant was probably not LoF or not LoF, with the exception of allele balance  $\leq 25\%$ . Mapping errors were often identified by an enrichment of complex variation surrounding a variant of interest. Furthermore, the UCSC browser was used to highlight mapping discrepancies, such

as self-chain alignments, segmental duplications, simple tandem repeats, and microsatellite regions.

In-frame multi-nucleotide variants (MNVs), essential splice site rescue, and frame-restoring insertion-deletions are rescue events that are predicted to restore gene function. MNVs were visualized in IGV and cross checked with codons from the UCSC browser; in frame MNVs that rescued stop codons were scored as not LoF. Essential splice site rescue occurs when an in frame alternative donor or acceptor site is present, which probably has a minimal effect on the transcript. A total of 36 base pairs upstream and downstream of the splice variant were assessed for splice site rescue. Cryptic splice sites within 6 base pairs of the splice variant were considered a complete rescue, rendering the variant not LoF. Rescue sites  $> 6$  base pairs away but within  $\pm 20$  base pairs were weighted with less confidence, scoring as probably not LoF. All potential splice site rescues were validated using Alamut v.2.11 (<https://www.interactive-biosoftware.com/alamut-visual/>). Frame-restoring indels were identified by scanning approximately  $\pm 80$  base pairs from the annotated indel and counting any insertions/deletions to assess if the frame would be restored.

Transcript errors encompass issues surrounding alternative transcripts, variants within a terminal coding exon, poorly conserved exons, and re-initiation events. Coding variants that occupied the minority ( $< 50\%$ ) of NCBI coding RefSeq transcripts for a given gene were considered not LoF. These variants often affected poorly conserved exons, as determined by PhyloP<sup>37</sup>, PhyloCSF<sup>19</sup> and visualization in the UCSC browser<sup>36</sup>. The only exceptions to the minority of transcript criteria were cases where the exon was well conserved, which relegated the categorization to probably not LoF. Variants within the last coding exon, or within 50 base pairs of the penultimate coding exon were also considered not LoF, unless  $25\% < x < 50\%$  of the coding sequence was affected, in which case the variant was deemed probably not LoF. If  $> 50\%$  of the coding sequence was disrupted by a variant in the last exon, this was deemed probably LoF. Other transcript errors included: re-initiation errors; upstream stop codons of a given LoF variant; variants that fell on exactly 50% of coding RefSeq transcripts; and/or partial exon conservation. Re-initiation events were flagged when a methionine downstream of the variant in the first coding exon was predicted to restart transcription, and were predicted to be probably not LoF. Variants occurring after a stop codon in the last coding exon were considered not LoF, particularly across the region of the exon or transcript in question. Error categories were grouped for Fig. 1 as follows: Minority of transcripts and weak exon conservation were grouped as transcript errors, genotyping errors and homopolymers as sequencing errors, essential splice rescue and MNV grouped as rescue and strand bias was included in other annotation errors.

The criteria above were strictly adhered throughout and manual curation was performed by two independent reviewers to ensure maximum consistency and minimize human error. Any discordance in curation was re-curated by both curators together and resolved. Full results of manual curation are available in Supplementary Table 3.

### Calculation of transcript-expression aware annotation

We first imported the GTEx v7 isoform quantifications into Hail and calculated the median expression of every transcript per tissue. This precomputed summary isoform expression matrix is available for GTEx v7 in [gs://gnomad-public/papers/2019-tx-annotation/data/GRCH37\\_hg19/](https://gnomad-public/papers/2019-tx-annotation/data/GRCH37_hg19/). We also import and annotate a variant file with the Variant Effect Predictor (VEP) version 85<sup>38</sup> against Gencode v19<sup>20</sup>, implemented in Hail with the LOFTEE v1.0 plugin.

We use the transcript consequences VEP field to calculate the sum of isoform expression for variant annotations, that is, the annotation-level expression across transcripts (ext). For variants that have multiple consequences for one transcript (for example, a single nucleotide variant

that is both a missense and a splice region variant on one transcript) we use the worst consequence, ordered by VEP (in this example, missense takes precedence over splice region). We filter the consequences to those only occurring on protein coding transcripts. Full ordering of the VEP consequences is available at: [useast.ensembl.org/info/genome/variation/prediction/predicted\\_data.html](http://useast.ensembl.org/info/genome/variation/prediction/predicted_data.html)

We then sum the expression of every transcript per variant, for every combination of consequence, LOFTEE filter, and LOFTEE flag for every tissue (Supplementary Fig. 3a). For example, if a single nucleotide variant is synonymous on ENST1, a high-confidence LOFTEE stop-gained variant on ENST3 and ENST4, and low-confidence LOFTEE stop-gained variant on ENST5 and ENST6, the ext values will be synonymous: ENST1, stop-gained high-confidence: ENST3 + ENST4, and stop-gained low-confidence: ENST5 + ENST6 per tissue. This can be computed with the `tx_annotate()` function by setting the `tx_annotation_type` to 'expression'. We foresee the non-normalized ext values to be useful when only considering one tissue of interest.

To allow for taking average expression values across tissues of interest, we normalize the expression value for a given value to the total expression of the gene on which the variant is found. This is carried out by dividing the ext value with the sum of the expression of all transcripts per tissue in transcripts per million (TPM) (Supplementary Fig. 3b). The resulting pext value can be interpreted as the proportion of the total transcriptional output from a gene that would be affected by the given variant annotation in question. If the gene expression value (and thus the denominator) in a given tissue is 0, the pext value will not be available (NA) for that tissue.

When taking averages across tissues, such unavailable pext values are not considered (that is, when taking the mean across tissues, we remove NA values). This value can be computed with the `tx_annotate()` function by setting the `tx_annotation_type` to 'proportion'. For the analyses in this manuscript, we remove reproduction-associated GTEx tissues (endocervix, ectocervix, fallopian tube, prostate, uterus, ovary, testes and vagina), cell lines (transformed fibroblasts and transformed lymphocytes) and any tissue with less than 100 samples (bladder, brain Cervicalc-1spinal cord, brain substantia nigra, kidney cortex and minor salivary gland), resulting in the use of 38 GTEx tissues.

We note that for a minority of genes, when RSEM<sup>15</sup> assigns higher relative expression to non-coding transcripts, the sum of the value of coding transcripts can be much smaller than the gene expression value for the transcript, resulting in low pext scores for all coding variants in the gene, and thus resulting in possible filtering of all variants for a given gene. In many cases this seems to be the result of spurious non-coding transcripts with a high degree of exon overlap with true coding transcripts. To prevent this artefact from affecting our analyses, we first calculated the maximum pext score for all variants across all protein-coding genes, and removed any gene where the maximum pext score was below 0.2. This resulted in the filtering of 668 genes, representing 3.3% of all genes analysed. We note that there is no overlap with the 668 genes and the haploinsufficient gene list, 97 of the filtered genes are present in OMIM (representing 1.5% of the OMIM gene list) and 42 filtered genes are considered constrained (representing 1.4% of LOEUF <0.35, or constrained, genes) thus having low effect on variant interpretation in the context of disease associations.

The full transcript-expression aware annotation pipeline, implemented in Hail 0.2, is fully available at [https://github.com/macarthur-lab/tx\\_annotation](https://github.com/macarthur-lab/tx_annotation) with commands laid out for analyses in the manuscript. Passing a Hail table through the `tx_annotate()` function returns the same table with a new field entitled 'tx\_annotation' which provides either the ext or pext value per variant-annotation pair, depending on parameter choice. We provide a helper function to extract the worst consequence and the associated expression values for these annotations. All analyses in the manuscript are based on the worst consequence of variant, ordered by VEP<sup>38</sup>.

## Functional validation of transcript-expression aware annotation

Conservation analysis was performed using phyloCSF scores using the same file used for the LOFTEE plugin, available publicly in [gs://gnomad-public/papers/2019-tx-annotation/data/other\\_data/phylocsf\\_data.tsv.bgz](https://gnomad-public/papers/2019-tx-annotation/data/other_data/phylocsf_data.tsv.bgz). We denoted exons with a phyloCSF max open-reading frame score >1,000 as highly conserved and those with phyloCSF max open-reading frame score <−100 as lowly conserved (Supplementary Fig. 5a) and evaluated their average usage in GTEx.

Using the base-level pext values that are used in the gnomAD browser, we filtered to intervals with high or low conservation, and calculated the average pext value in the interval. To evaluate regions with low conservation but high expression, we identified genes harbouring unconserved regions with the pext value >0.9 for pathway enrichment analysis and used the web browser for FUMA GENE2FUNC feature<sup>39</sup>, which incorporates Reactome<sup>40</sup>, KEGG<sup>41</sup>, Gene Ontology<sup>42</sup> (GO) as well as other ontologies. Default parameters were used for FUMA, with all protein coding genes as the background list. Results from FUMA pathway analysis are available in Supplementary Fig. 12, and full results are available in Supplementary Table 7.

Analysis of pext values for LOFTEE flags and the MAPS calculation were performed using the gnomAD v2.1.1 exome dataset. Calculation of MAPS scores was previously described<sup>21</sup> and is implemented as a Hail module, as also described previously<sup>1</sup>. MAPS is a relative metric, and cannot be compared across datasets, but is a useful summary metric for the frequency spectrum, indicating deleteriousness as inferred from rarity of variation (high values of MAPS correspond to lower frequency, suggesting the action of negative selection at more deleterious sites). The MAPS scores were calculated on the gnomAD v2.1.1 dataset partitioning upon the LOEUF score and expression bin. The script for generating MAPS scores is available in the tx-annotation Github repository under `/analyses/maps/maps_submit_per_class.py`

## Manual evaluation of unexpressed regions in haploinsufficient developmental delay genes using the GENCODE workflow

As an orthogonal evaluation of regions flagged as unexpressed with the pext metric, we identified any region in 61 haploinsufficient disease genes with a mean pext value <0.1 in all GTEx tissues and in GTEx brain samples, owing to the relevance of brain tissues for these disorders, regardless of mutational burden in gnomAD. The resulting list of 128 regions was evaluated by the HAVANA manual annotation group of the GENCODE project<sup>20</sup>.

The manual evaluation first established whether the transcript model corresponding to the region in question was correct in terms of structure, comparing exon–intron combinations, and the accuracy of splice sites against the RNA evidence supporting the model. Second, the functional biotype of each model was reassessed; in particular, whether the decision to annotate the model as protein-coding in GENCODE v19 was appropriate. Note that GENCODE models that incorporate alternative exons or exon combinations in comparison to the 'canonical' isoform are likely to be annotated as coding if they contain a prospective CDS that is considered biologically plausible, based on a mechanistic view of translation. These re-annotations are summarized in Supplementary Table 5.

We binned cases into three main categories, according to confidence in both the accuracy and potential functional relevance of the overlapping models: (1) 'error', in which the model was seen to have an incorrect transcript structure and/or a CDS that conflicted with updated GENCODE annotation criteria (these annotations had been or will be changed in future GENCODE releases based on this evaluation); (2) 'putative', in which the model structure and CDS satisfied our current annotation criteria, although we judged the potential of the transcript represented to encode a protein with a functional role in cellular physiology to be nonetheless speculative (these have

been maintained as putative protein-coding transcripts in GENCODE); (3) ‘validated’, in which we believe it is highly probable that the model represents a true protein-coding isoform. High confidence in the validity of the CDS was based on comparative annotation, that is, the observation of CDS conservation and also the existence of equivalent transcript models in other species. GENCODE also annotates transcript models as ‘nonsense-mediated decay’ and ‘non-stop decay’, in which a translation is found that is predicted to direct the RNA molecule into cellular degradation programs. Although it has been established that such ‘non-productive’ transcription events can have a role in gene regulation and thus disease, the interpretation of variants within nonsense-mediated decay and non-stop decay CDS regions remains challenging<sup>43</sup>. These models were therefore classed in a separate category.

### Gene list comparisons

To evaluate the filtering power of the pext metric for Mendelian variants, we evaluated the number of variants that would be filtered with an average GTEx pext cutoff of 0.1 (low expression) in the ClinVar and gnomAD datasets. We downloaded the ClinVar VCF from the ClinVar FTP (version dated 10/28/2018), imported it into Hail, annotated it with VEP v85 against Gencode v19, and added pext annotations with the tx\_annotate() function. All evaluated variants were annotated as HC by LOFTEE v1.0, and ClinVar variants were filtered to those marked as pathogenic, with no conflicts, and reviewed with at least one star status.

For variants in 61 haploinsufficient genes, we identified any variant identified in at least one individual with any zygosity in both datasets. For variants identified in autosomal recessive disease genes, we used a list of 1,183 OMIM disease genes deemed to follow a recessive inheritance pattern by Blekhman et al.<sup>44</sup> and Berg et al.<sup>45</sup> (available as [https://github.com/macarthur-lab/gene\\_lists/blob/master/lists/all\\_ar.tsv](https://github.com/macarthur-lab/gene_lists/blob/master/lists/all_ar.tsv)). We compared the pext value for all pLoF variants identified in ClinVar versus any variant in a homozygous state in at least one individual in the gnomAD exome or genome datasets. Finally, we used a LOEUF cutoff of 0.35 to denote constrained genes, and compared any synonymous or pLoF variant in these genes in the gnomAD exome or genome datasets.

### De novo and rare variant analysis

De novo variants were collated from previously published studies. We collected de novo variants identified in 5,305 probands from trio studies of intellectual disability/developmental disorders (Hamdam et al.<sup>27</sup>:  $n = 41$ , de Ligt et al.<sup>28</sup>:  $N = 100$ , Rauch et al.<sup>29</sup>:  $N = 51$ , DDD<sup>24</sup>:  $n = 4,293$ , Lelieveld et al.<sup>26</sup>:  $n = 820$ ), 1,073 probands with congenital heart disease with co-morbid developmental delay (Sifrim et al.<sup>46</sup>:  $n = 512$ , Chih Jin et al.<sup>47</sup>: 561), 6,430 ASD probands, and 2,179 unaffected controls from the Autism Sequencing Consortium<sup>25</sup>. We also used a previously published dataset of variants in 8,437 cases with ASD and/or attention-deficit/hyperactivity disorder and 5,214 controls from the Danish Neonatal Screening Biobank<sup>48</sup>. In this analysis, we analysed pLoF variants identified in highly constrained genes (first LOEUF decile) with a combined total allele count of  $\leq 10$  in cases and controls.

We annotated both de novo and rare variants with VEP v85 against Gencode v19 and added pext annotations with the tx\_annotate() function. We then calculated the average pext metric across 11 GTEx brain samples and binned them as low ( $\text{pext} < 0.1$ ), medium ( $0.1 \leq \text{pext} \leq 0.9$ ) or high ( $\text{pext} > 0.9$ ) expression. We then calculated the number of pLoF, missense, and synonymous variants per pext expression bin. To obtain case-control rate ratios and the 95% confidence intervals for de novo variant analyses, we used a two-sided Poisson exact test on counts. To obtain the odds ratio for the rare variant analysis in ASD/ADHD, we used the Fisher’s exact test for count data.

### Isoform quantifications via salmon

To evaluate whether use of a different isoform quantification tool would affect results, we compared results of *TCF4* base-level expression

(shown in Fig. 2b), MAPS (Fig. 3c) and comparison of the number of variants filtered in haploinsufficient developmental disease genes in ClinVar vs gnomAD (Fig. 4a) using RSEM quantifications used in this study with quantifications using salmon v.0.12<sup>47</sup>. Due to the intractability of re-quantifying the entire GTEx dataset, we downloaded and requantified 151 GTEx brain cortex CRAM files from the V7 dataset. We first converted CRAMs to fastq files using Picard 2.18.20 and ran salmon with the ‘salmon quant -i index -fastq1 - fastq2 -minAssigned-Frag1 -validateMappings’ command. The index was created with the ‘salmon index -t transcript.fa -type quasi -k 31’ command using the GENCODE v19 protein-coding and lncRNA transcripts FASTA files. The existing GTEx RSEM isoform quantifications were filtered to the same GTEx brain cortex samples. For the analyses to remain consistent with the remainder of the manuscript, we calculated the maximum brain cortex pext score for all variants across all protein-coding genes for both the RSEM and salmon quantifications, and removed any gene in which the maximum pext score was below 0.2. This resulted in filtering 325 genes from the salmon quantification of the brain cortex samples and 691 genes from the RSEM quantification, corresponding to 3.4 and 1.6% of quantified genes, respectively. We filtered these genes in both the MAPs and gene list comparison analysis seen in Supplementary Fig. 11. The WDL script for the quantification pipeline is available at: [gs://gnomad-public/papers/2019-tx-annotation/results/salmon\\_rsem/salmon.wdl](https://gnomad-public/papers/2019-tx-annotation/results/salmon_rsem/salmon.wdl) and the commands to obtain results for each individual analysis in the tx-annotation Github repository under /analyses/rsem\_salmon/.

### Transcript expression aware annotation with fetal isoform expression dataset

Although our analyses were based on transcript expression aware annotation from the GTEx v7 dataset, we provide necessary files for pext annotation with the Human Brain Development Resource (HBDR) fetal brain dataset<sup>49</sup> in [gs://gnomad-public/papers/2019-tx-annotation/data/HBDR\\_fetal\\_RNaseq](https://gnomad-public/papers/2019-tx-annotation/data/HBDR_fetal_RNaseq). HBDR includes 558 samples from varying brain subregions across developmental time points. We downloaded HBDR sample fastq files from European Nucleotide Archive (study accession PRJEB14594) and obtained RSEM isoform quantification on HBDR fastqs using the GTEx v7 quantification pipeline, publicly available at <https://github.com/broadinstitute/gtex-pipeline/> which briefly involves two-pass alignment with STAR v2.4.2a<sup>50</sup> and isoform quantification with RSEM v1.2.22. Here, we also removed genes where the average pext across HBDR was below 0.2, resulting in the removal of 712 genes (3.5% of all analysed genes). The dataset was also used for the analysis of baselevel expression values in *SCN2A* shown in Supplementary Fig. 7d.

### Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

### Data availability

We used the gnomAD v.2.1.1 sites Hail 0.2 (<https://hail.is>) table that is accessible publicly at [gs://gnomad-public/release/2.1.1](https://gnomad-public/release/2.1.1) and at <https://gnomad.broadinstitute.org>. The GTEx v7 gene and isoform expression data were downloaded from the GTEx portal ([gtexportal.org](https://gtexportal.org)). The LOEUF constraint file was downloaded from [gs://gnomad-resources/lof\\_paper/](https://gnomad-resources/lof_paper/). All files used in the analyses in the manuscript are available in [gs://gnomad-public/papers/2019-tx-annotation/](https://gnomad-public/papers/2019-tx-annotation/).

### Code availability

The GTEx pipeline for isoform quantification is available publicly (<https://github.com/broadinstitute/gtex-pipeline/>) and briefly involves two-pass alignment with STAR v2.4.2a<sup>50</sup>, gene expression quantification

with RNA-SeQC v1.1.8<sup>51</sup>, and isoform quantification with RSEM v1.2.22. Variants used in all gnomAD analyses in the manuscript passed random forest filtering, and all pLoF variants were annotated as high-confidence by LOFTEE v1.1.0, which is described in an accompanying manuscript<sup>1</sup>. Scripts to quality control the gnomAD dataset are available at [https://github.com/macarthur-lab/gnomad\\_qc](https://github.com/macarthur-lab/gnomad_qc) and the scripts to generate files for the analyses are available at [https://github.com/macarthur-lab/tx\\_annotation](https://github.com/macarthur-lab/tx_annotation).

34. Riggs, E. R. et al. Copy number variant discrepancy resolution using the ClinGen dosage sensitivity map results in updated clinical interpretations in ClinVar. *Hum. Mutat.* **39**, 1650–1659 (2018).
35. Thorvaldsdóttir, H., Robinson, J. T. & Mesirov, J. P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief. Bioinform.* **14**, 178–192 (2013).
36. Kent, W. J. et al. The Human Genome Browser at UCSC. *Genome Res.* **12**, 996–1006 (2002).
37. Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R. & Siepel, A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.* **20**, 110–121 (2010).
38. McLaren, W. et al. The Ensembl Variant Effect Predictor. *Genome Biol.* **17**, 122 (2016).
39. Watanabe, K., Taskesen, E., van Bochoven, A. & Posthuma, D. Functional mapping and annotation of genetic associations with FUMA. *Nat. Commun.* **8**, 1826 (2017).
40. Croft, D. et al. Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Res.* **39**, D691–D697 (2011).
41. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
42. The Gene Ontology Consortium. Gene Ontology: tool for the unification of biology. *Nat. Genet.* **25**, 25–29 (2000).
43. da Costa, P. J., Menezes, J. & Romão, L. The role of alternative splicing coupled to nonsense-mediated mRNA decay in human disease. *Int. J. Biochem. Cell Biol.* **91** (Pt B), 168–175 (2017).
44. Blekman, R. et al. Natural selection on genes that underlie human disease susceptibility. *Curr. Biol.* **18**, 883–889 (2008).
45. Berg, J. S. et al. An informatics approach to analyzing the incidentalome. *Genet. Med.* **15**, 36–44 (2013).
46. Sifrim, A. et al. Distinct genetic architectures for syndromic and nonsyndromic congenital heart defects identified by exome sequencing. *Nat. Genet.* **48**, 1060–1065 (2016).
47. Jin, S. C. et al. Contribution of rare inherited and de novo variants in 2,871 congenital heart disease probands. *Nat. Genet.* **49**, 1593–1601 (2017).
48. Satterstrom, F. K. et al. ASD and ADHD have a similar burden of rare protein-truncating variants. Preprint at <https://www.biorxiv.org/content/10.1101/277707v1> (2018).
49. Lindsay, S. J. et al. HDBR expression: a unique resource for global and individual gene expression studies during early human brain development. *Brain Dev.* **10**, 86 (2016).

50. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
51. DeLuca, D. S. et al. RNA-SeQC: RNA-seq metrics for quality control and process optimization. *Bioinformatics* **28**, 1530–1532 (2012).

**Acknowledgements** We thank all of the research participants for contributing their data. This work was supported by NIDDK U54 DK105566, NIGMS R01 GM104371, and the Broad Institute. KJK was supported by NIGMS F32 GM115208. A.O.L. was supported by NICHD K12 HD052896. The GENCODE project is supported by the National Human Genome Research Institute of the National Institutes of Health under Award Number U41HG007234. The results published here are in part based on data: (1) generated by The Cancer Genome Atlas (TCGA) managed by the NCI and NHGRI (accession: phs000178.v10.p8); information about TCGA can be found at <http://cancergenome.nih.gov>; (2) generated by the Genotype-Tissue Expression Project (GTEx) managed by the NIH Common Fund and NHGRI (accession: phs000424.v7.p2); (3) generated by the Exome Sequencing Project, managed by NHLBI; and (4) generated by the Alzheimer's Disease Sequencing Project (ADSP), managed by the NIA and NHGRI (accession: phs000572.v7.p4). We thank E. Pierce-Hoffman for previous analysis and thoughts on characterizing loss-of-function variants in haploinsufficient genes. We thank the iPSYCH/SSI/Broad Institute psychiatric genetics study for the use of exome count data. We have complied with all relevant ethical regulations. This study was overseen by the Broad Institute's Office of Research Subject Protection and the Partners Human Research Committee, and was given a determination of Not Human Subjects Research. Informed consent was obtained from all participants. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

**Author contributions** B.B.C., K.J.K., J.A. and D.G.M. contributed to the writing of the manuscript and generation of figures. B.B.C., K.J.K., E.G.S., M.S.-B., J.M.M., J.A.K., M.J.D. and D.G.M. contributed to the analysis of data. B.B.C., K.J.K., J.A.K., J.M.M., J.K., F.K.S., A.H.O'D.-L., T.P., C.S., M.J.D. and D.G.M. developed tools and methods that enabled the scientific discoveries herein. B.B.C., K.J.K., N.A.W., T.P., M.S., J.A., M.J.D. and D.G.M. contributed to the production and quality control of the gnomAD dataset. All authors listed under The Genome Aggregation Database Consortium contributed to the generation of the primary data incorporated into the gnomAD resource. All authors reviewed the manuscript.

**Competing interests** K.J.K. owns stock in Personalis. A.H.O'D.-L. has received honoraria from ARUP and Chan Zuckerberg Initiative. M.J.D. is a founder of Maze Therapeutics. D.G.M. is a founder with equity in Goldfinch Bio, and has received research support from AbbVie, Astellas, Biogen, BioMarin, Eisai, Merck, Pfizer, and Sanofi-Genzyme.

## Additional information

**Supplementary information** is available for this paper at <https://doi.org/10.1038/s41586-020-2329-2>.

**Correspondence and requests for materials** should be addressed to D.G.M.

**Peer review information** *Nature* thanks Roderic Guigo and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

**Reprints and permissions information** is available at <http://www.nature.com/reprints>.

## Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see [Authors & Referees](#) and the [Editorial Policy Checklist](#).

### Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

- |                                     |  |
|-------------------------------------|--|
| n/a                                 | Confirmed  |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> The exact sample size ( $n$ ) for each experimental group/condition, given as a discrete number and unit of measurement  |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly  |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> The statistical test(s) used AND whether they are one- or two-sided<br><i>Only common tests should be described solely by name; describe more complex techniques in the Methods section.</i>   |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> A description of all covariates tested  |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons  |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals) |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> For null hypothesis testing, the test statistic (e.g. $F$ , $t$ , $r$ ) with confidence intervals, effect sizes, degrees of freedom and $P$ value noted<br><i>Give <math>P</math> values as exact values whenever suitable.</i>                            |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings  |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes  |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> Estimates of effect sizes (e.g. Cohen's $d$ , Pearson's $r$ ), indicating how they were calculated   |

*Our web collection on [statistics for biologists](#) contains articles on many of the points above.*

### Software and code

Policy information about [availability of computer code](#)

Data collection No software was used for the collection of data, as this was an opportunistic study.

Data analysis All code to perform quality control and data analysis is provided in the following Github repos:  
[https://github.com/macarthur-lab/tx\\_annotation](https://github.com/macarthur-lab/tx_annotation)  
[https://github.com/macarthur-lab/gnomad\\_qc](https://github.com/macarthur-lab/gnomad_qc)  
[https://github.com/macarthur-lab/gnomad\\_lof](https://github.com/macarthur-lab/gnomad_lof)  
<https://github.com/konradjk/loftree>  
<https://github.com/broadinstitute/gtex-pipeline/>  
Hail 0.2 is available at: <https://hail.is/>  
RSEM v1.2.22 : <https://deweylab.github.io/RSEM/>  
RNA-SeQC v1.1.8 : <https://github.com/broadinstitute/rnaseqc>  
STAR v2.4.2a : <https://github.com/alexdobin/STAR>  
Alamut v.2.11 <https://www.interactive-biosoftware.com/alamut-visual/>  
Variant Effect Predictor (VEP) v.85 : <https://uswest.ensembl.org/info/docs/tools/vep/index.html>  
FUMA GENE2FUNC v1.3.5e : <https://fuma.ctglab.nl/>  
R version 3.4.0 : <https://cran.r-project.org/bin/macosx/>

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research [guidelines for submitting code & software](#) for further information.



## Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All datasets are described in the manuscript or Supplementary Information, including deposition of the full dataset at <https://gnomad.broadinstitute.org>. Data for specific analyses are available publicly at [gs://gnomad-public/papers/2019-tx-annotation/](https://gnomad-public/papers/2019-tx-annotation/) and the specific folders therein for analyses are referenced in the manuscript for ease of recreating analyses with the data provided. There are no restrictions on the aggregate data released.

## Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

## Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	This study provides a framework and tool to improve variant interpretation in datasets of any size, no matter how small or large. As a proof of principle, we use one of the largest datasets of human genetic variation, gnomAD and the largest functional genomics dataset GTEx. In other words, this study is opportunistic, and involves secondary use of available genome, exome and transcriptome data. No sample size was predetermined.
Data exclusions	Sample QC and variant QC for gnomAD are described extensively in the supplementary methods of the main manuscript. Notably, individuals with severe pediatric disease, and known first disease relatives of those with severe pediatric disease were excluded. For the analyses in the manuscript, we removed GTEx tissues with low sample numbers, reproductive tissue and non-tissues (ie. cell lines). For the purpose of our manuscript, we did not define pre-exclusion criteria for calculation of pext. However for the analyses in the manuscript, we defined pre-exclusion tissues : we removed reproduction-associated GTEx tissues (endocervix, ectocervix, fallopian tube, prostate, uterus, ovary, testes, vagina), cell lines (transformed fibroblasts, transformed lymphocytes) and any tissue with less than one hundred samples (bladder, brain Cervicalc-1 spinal cord, brain substantia nigra, kidney cortex, minor salivary gland) . This is explained in the Methods section of the manuscript.
Replication	We did not attempt to reproduce any findings in a separate but identical dataset, as no other data set of comparable size exists. However we replicate key findings with a separate isoform quantification tool. We successfully replicate the MAPS results, shown in Supplementary Figure 11. We also use an external fetal dataset to provide additional data.
Randomization	As this was a population-based study, and not a case-control study, no randomization was performed.
Blinding	As this was a population-based study, and not a case-control study, blinding was not relevant.

## Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

### Materials & experimental systems

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> Antibodies
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms
<input type="checkbox"/>	<input checked="" type="checkbox"/> Human research participants
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data

### Methods

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging

# Human research participants

Policy information about [studies involving human research participants](#)

Population characteristics	<p>As an opportunistic collection of data, the participants in this study were not selected based on age, gender, or genotypic information. As described above, individuals with severe pediatric disease, and known first disease relatives of those with severe pediatric disease were excluded from gnomAD. As an opportunistic collection of data, the participants in gnomAD were not selected based on age, gender, or genotypic information. The populations are provided in Supplementary Table 7 of the accompanying Karczewski et al., and there are 64,754 females and 76,702 males. These data were obtained primarily from case-control studies of adult-onset common diseases, including cardiovascular disease, type 2 diabetes, and psychiatric disorders.</p> <p>GTEx v7 collection was similarly opportunistic and has been previously extensively published and reported on. Population characteristics of the data can be found in Reference 2 : GTEx Consortium et al., Genetic effects on gene expression across human tissues. Nature 550, 204 (2017).</p>
Recruitment	<p>As this was an opportunistic secondary use study, we did not recruit any participants.</p>
Ethics oversight	<p>This study was overseen by the Broad Institute’s Office of Research Subject Protection and the Partners Human Research Committee, and was given a determination of Not Human Subjects Research.</p>

Note that full information on the approval of the study protocol must also be provided in the manuscript.